BioE/MCB/PMB C146/246, Spring 2003

Problem Set 5: Biological Significance of Alignments; Multiple Alignment

Due 26 Feb 03, 5:00 pm PST by email to derek@rana.lbl.gov

1. 5 points
The best substitution matrix for Smith-Waterman comparisons of distant homologs is often BLOSUM45. Which BLOSUM matrices would you use for BLAST comparisons of distant homologs? Why?

2. 5 points
When will an optimal alignment not be found by FASTA? By BLAST?

3. 5 points
Why is it necessary to use masking for low complexity regions? Why is it necessary to use masking for coiled coil regions?

4. 5 points
If looking for similar protein-coding regions in two unannotated genome (nucleotide) sequences, what BLAST program would you use? Why?

5. 5 points
Name two features of genomes that are not protein coding regions. What BLAST programs would you use to find these similar features in two unannotated genome sequences? Why?

6. 10 points

(A) What are sources of errors in functional annotations of protein sequences?

(B) A BLASTP search was conducted on a hypothetical open reading frame. Given the BLAST results and Gene Ontology assignments for the top hits on the next page, what can you deduce about the functional roles of the unannotated query protein?

WU-BLAST RESULTS

High Score	P_N	# HSP	Biological Process	Cellular Component	Molecular Function	Evidence
229	4.60E-17	9	Hydrogen transport	unclassified	ATP binding, Ca^{2+} binding	Inferred by electronic annotation
140	9.30E-17	8	Intracellular signaling cascade	unclassified	unclassified	Inferred by electronic annotation
172	3.30E-11	4	Intracellular signaling cascade	unclassified	ATP binding, protein tyrosine kinase	Inferred by electronic annotation
162	9.00E-11	4	Intracellular signaling cascade, programmed cell death	plasma membrane	SH3/SH2 adaptor protein	Inferred by mutant phenotype & sequence similarity
126	2.70E-10	8	Unclassified	unclassified	unclassified	
164	9.10E-10	2	Intracellular signaling cascade	unclassified	ATP binding, protein tyrosine kinase	Inferred by electronic annotation
160	1.20E-09	1	Intracellular signaling cascade	unclassified	ATP binding, protein tyrosine kinase	Inferred by electronic annotation
154	3.80E-09	3	Unclassified	unclassified	unclassified	
121	4.60E-09	6	Unclassified	unclassified	unclassified	
122	1.80E-06	3	Unclassified	unclassified	unclassified	
130	2.20E-06	4	protein amino acid phosphorylation	unclassified	ATP binding, protein serine/threonine kinase	Inferred by electronic annotation
96	2.40E-06	7	Unclassified	unclassified	unclassified	
119	4.20E-06	3	Unclassified	unclassified	unclassified	
124	1.00E-05	2	Unclassified	unclassified	unclassified	

7. 10 points

Infer the functions of the unknown genes, given the phylogenetic profiles of their orthologs in various species. How confident are you of these predictions?

Species Abbreviations

E	*F*	*G*	*H*	*J*	*R*	*S*	*Name*	*Function*
1	1	1	0	1	1	1	CheZ	Chemotaxis
1	1	1	0	1	1	1	CheY	Chemotaxis
1	1	1	1	1	0	1	EnvZ	Histidine kinase
1	0	1	1	0	0	0	FimA	Type I Pilin
1	0	1	1	0	0	0	FimG	Type I Pilin
1	0	1	0	1	0	0	PilA	Type IV Pilin
1	1	1	0	1	1	1	Unknown 1	
1	0	1	1	1	0	0	Unknown 2	

8. 10 points

Infer the functions of the unknown protein-coding genes, using a domain fusion approach:

Domain Abbreviations

A	*B*	*C*	*D*	*E*	*Species*	*Name*	*Function*
0	0	1	0	0	*A. nidulans*	CPSase	Carbamoyl-phosphate synthetase
0	0	0	1	0	*A. nidulans*	ATCase	Aspartate transcarbamylase
0	0	0	0	1	*A. nidulans*	DHOase	Dihydroorotase
1	0	0	0	0	*E. coli*	TrpC	Tryptophan biosynthesis
0	1	0	0	0	*E. coli*	TrpG	Tryptophan biosynthesis
1	1	0	0	0	*S. cerevisiae*	Unknown 1	
1	0	1	1	1	*D. melanogaster*	Unknown 2	

9. 25 points

(A) The following protein coding sequences are available from the course website as `c246_2003_ps5_seq.fasta`. Download MSA from the NCBI website (http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html) and align the sequences:

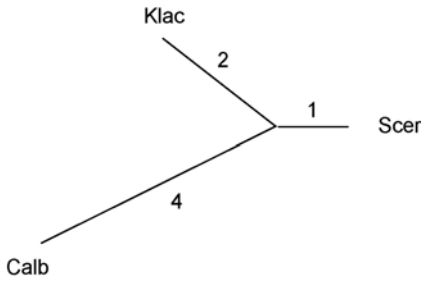
```
>Scer_Cbf1
ATTDEWKKQRKDSHKEVERRRRENINTAINVLSDLLPVRESSKAAILACAAEYIQKLKET
DEANIEKWTLQKLLSEQNASQLASANEKLQEELGNAYKEIEYMKRVLRK

>Calb_Cbf1
HGSEEWHRQRRENHKEVERKRRESINTGIRELARLIPTTDTNKAQILQRAVEYIKRLKEN
ENNNIEKWTLEKLLTEQAVSELSASNEKLKHELESAYREIEQLKRGKK

>Klac_Cbf1
TGSTAWKQQRKESHKEVERRRRQNINTAIEKLSDLLPVKETSKAAILSRAAEYIQKMKET
ETANIEKWTLQKLLGEQQVSSLTSANDKLEQELSKAYKNLQELKKKLKEAGIEDPTEEE
```

(B) Calculate the minimum entropy score for this multiple alignment

(C) Calculate the sum of pairs score for this multiple alignment

(D) Calculate the star tree distance for this multiple alignment. Use a distance metric of +1 for a match and –1 for a mismatch

(E) Calculate the tree distance for this multiple alignment, using the following tree:



10. 20 points

(A) Download the following members of the `Arthro_defensin` protein family from PFAM: DEFI_APIME/53-82, DEFI_AESCY/1-37; DEFA_ZOPAT/10-43, SAPC_SARPE/10-39. Perform pairwise alignments (global alignment with no end gap penalties) between the first sequence and each of the other sequences, and assemble a master-slave alignment. Compare your alignment with the Pfam alignment.

(B) Obtain alignments of these protein domains from TWO other databases. Compare and contrast the three alignments (Pfam and two other database alignments).

Extra credit (5 points)
Given the output from 5 different alignment algorithms, how would you determine the best alignment? (*Hint*: Using methods similar to question 9 do not provide sufficiently independent means for validation.)